DOCUMENT RESUME

ED 431 794 TM 029 880

AUTHOR Tirri, Henry; Silander, Tomi
TITLE Stochastic Complexity Based Estimation of Missing Elements in Questionnaire Data.
PUB DATE 1998-04-00
NOTE 12p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Elementary Education; Elementary School Students; *Estimation (Mathematics); Foreign Countries; *Information Theory; *Questionnaires; *Research Methodology
IDENTIFIERS *Missing Data; *Stochastic Analysis

ABSTRACT

A new information-theoretically justified approach to missing data estimation for multivariate categorical data was studied. The approach is a model-based imputation procedure relative to a model class (i.e., a functional form for the probability distribution of the complete data matrix), which in this case is the set of multinomial models with some independence assumptions. Based on the given model class assumption, an information-theoretic criterion can be derived to select between the different complete data matrices. Intuitively this general criterion, called stochastic complexity, represents the shortest code length needed for coding the complete data matrix relative to the model class chosen. Using these information-theoretic criteria, the missing data problem is reduced to a search problem, that of finding the data completion with minimal stochastic complexity. The results of two empirical studies of the approach, using educational data sets of 478 elementary school students ("Popular kids" - POPKIDS in Michigan) and 500 Irish schoolchildren ("Irish educational transitions data - Irish), are presented and compared to those achieved with commonly used techniques such as case deletion and imputation of sample averages. (Contains 3 figures, 6 tables, and 36 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 431 794

Stochastic Complexity Based Estimation of Missing Elements in Questionnaire Data

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Henry Tirri
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Henry Tirri
Tomi Silander

This paper is prepared for the:
Annual Meeting of the American Educational Research Association in San Diego, CA
April 1998

TM029880

BEST COPY AVAILABLE

Stochastic Complexity Based Estimation of Missing Elements in Questionnaire Data

Henry Tirri
Complex Systems Computation Group (CoSCo)
P.O.Box 26, Department of Computer Science
FIN-00014 University of Helsinki, Finland

Tomi Silander
Complex Systems Computation Group (CoSCo)
P.O.Box 26, Department of Computer Science
FIN-00014 University of Helsinki, Finland

In this paper we study a new information-theoretically justified approach to missing data estimation for multivariate categorical data. The approach discussed is a model-based imputation procedure relative to a model class (i.e., a functional form for the probability distribution of the complete data matrix), which in our case is the set of multinomial models with some independence assumptions. Based on the given model class assumption an information-theoretic criterion can be derived to select between the different complete data matrices. Intuitively this general criterion, called stochastic complexity, represents the shortest code length needed for coding the complete data matrix relative to the model class chosen. Using this information-theoretic criteria, the missing data problem is reduced to a search problem, i.e., finding the data completion with minimal stochastic complexity. In the experimental part of the paper we present empirical results of the approach using two real data sets, and compare these results to those achived by commonly used techniques such as case deletion and imputating sample averages.

Introduction

In most educational research contexts the available data are typically incomplete and contain usually several elements with missing information. Missing elements are particularly typical to data from complex questionnaire based surveys, where the lack of time or low motivation of the respondents result in neglection of many of the questions. In most cases omission of incomplete data, i.e., concentration on only records with complete data, is infeasible, as the amount of data for the analysis would be drastically reduced. Therefore intelligent methods for handling missing data are an important aspect of quantitative data analysis.

The problem of missing data estimation has been addressed widely in the statistics literature (see e.g., (Gelman, Carlin, Stern, & Rubin, 1995; Rubin, 1987, 1996; Schafer, 1995)). The last quarter of a century has seen many developments in this area. The EM algorithm together with its extensions (Dempster, Laird, & Rubin, 1977; McLachlan & Thriyambakam, 1997), multiple imputation (Rubin, 1987, 1996; Schafer, 1995) and Markov Chain Monte Carlo (Gilks, Richardson, & J., 1996) all provide tools for inference in large classes of missing data problems. In practice, however, these developments have not had large impact on the way most data analysts handle missing values on a routine basis. This is partly due to the rather complex nature of these approaches, but mostly because of their lack of support in statistical software.

The purpose of this paper is to study a new information-theoretically justified approach to missing data estimation. The method discussed is deeply related to Bayesian inference, but originates from the research on universal coding (Rissanen, 1984), which aims at finding good (short) encodings of data. We do not make an attempt to provide a survey of the aforementioned developments in missing data estimation—an interested reader can consult the excellent books by Little and Rubin (1987) and Schafer (1997). In order to put our work in perspective, however, we would like to remind that the proposed methods can essentially be categorized into two general approaches: case deletion and imputation (Little & Rubin, 1987; Schafer, 1997). In *case deletion* all the cases with missing data are omitted and the analysis is performed only using the complete cases. Obviously this approach is a reasonable solution only if the incomplete cases comprise a small fraction of all cases. In *imputation-based procedures* the missing data values are filled with plausible values which forces the incomplete data set into a complete data format. The methods in this group vary from simple sample average imputation approaches (Little & Rubin, 1987) to complex multiple imputation procedures (Schafer, 1997). The latter share the same underlying philosophy as EM and data augmentation: an incomplete-data problem is solved by repeatedly solving the complete-data version. In multiple imputation the unknown missing data are replaced by several "simulated" values using Monte Carlo approaches, and each of the resulting complete data sets is analyzed by standard complete data methods. The resulting

This research has been supported by the Technology Development Center (TEKES), and by the Academy of Finland.

BEST COPY AVAILABLE

variability is then taken to reflect the uncertainty caused by the missing data.

The approach discussed here can be characterized as a model-based imputation procedure. Of the existing approaches, the method described here is somewhat related to Bayesianly proper multiple imputation (Schafer, 1997), which uses independent realizations of the posterior predictive distribution of the missing data under some complete-data model and prior. The method discussed here is similar in the sense that it is always relative to a model class, and the criterion used for finding the values to be imputed can be approximated by the Bayesian marginal likelihood (Berger, 1985; Bernardo & Smith, 1994; O'Hagan, 1994). However, we are interested in the problem of finding a *single optimal completion* of the incomplete data set instead of a set of completions typical to multiple imputation procedures. Moreover, as we will see in the next section, the augmentation criterion has its foundations in information and coding theory (Cover & Thomas, 1991) rather than Bayesian statistics.

Intuitively the approach can be described as follows. By modeling the set of data records as a matrix of incomplete discrete data, the missing part is estimated by assuming a functional form for the probability distribution of the data cases (i.e., a model class). Based on the given model class assumption an information-theoretic criteria can be derived to select between the different complete data matrices for the more "likely" one (in abstract sense). Intuitively this general criteria, called *stochastic complexity* (Rissanen, 1987, 1989, 1996) represents the shortest code length needed for coding the complete data matrix relative to the model class chosen. Unfortunately in general the exact criteria is very hard to compute for many interesting model families, but it can be approximated by the Bayesian marginal likelihood computed by integrating over all the possible models (parameter settings) in the chosen model class.

Since we are interested in categorical data, we use the set of (saturated) multinomial models for the complete data, which is a more general descriptive "language" than the multivariate normal class in the sense that it allows also for higher than two-way associations among the variables. In addition to selecting the multinomial model class additional independence assumptions for the variables have to be made to make the approach feasible in practice. This leads us to consider a special subclass of finite mixture models (Titterington, Smith, & Makov, 1985) known as Naive Bayes models.

Having defined an evaluation criterion for our data completions, the missing data problem is reduced to a search problem, where the goal is to find a data completion that minimizes the stochastic complexity of the completed matrix. Due to the large discrete search space exhaustive search for the minimal stochastic complexity completion among all the possible completions is not feasible for data sets with large fractions of missing data. However, locally optimal solutions can be found by using stochastic search methods such as EM and simulated annealing (Aarts & Korst, 1989). In this paper for the experiments we use a simple easy-to-implement variant called stochastic greedy, which has a comparable performance to the more complex search methods (Kontkanen, Myllymäki, Silander, & Tirri, 1997a).

In the experimental part of the paper we present empirical results of the approach using two real data sets, and compare these results to those achieved by commonly used techniques such as case deletion and imputating sample averages. It should be observed that albeit we discuss the finite mixture model class, the approach presented is general and applicable to imputation of categorical data with other model classes also.

The missing data problem

We will consider rectangular data sets whose rows can be modeled as independent, identically distributed (iid) draws from some multivariate probability distribution. The rows represent observational units and the columns represent variables recorded for those units. In the following such rows of the matrix are called *data vectors* $\vec{d}$, and data set D is defined as a set of N data vectors $\vec{d}_1, \ldots, \vec{d}_N$. Each complete data vector $\vec{d}$ consists of $m+1$ value assignments, $\vec{d} = (X_1 = x_1, \ldots, X_{m+1} = x_{m+1})$, where each value x_i is assumed to belong to the discrete set $\{x_{i1}, \ldots, x_{in_i}\}$. Consequently, a complete data set D can be regarded as a $N \times m+1$ matrix, where each component $\vec{d}_{ji}$ is a value assignment of the form $< X_i = x_i >$. In the incomplete data case, one or more of these assignments are initially unknown. In the sequel we partition the matrix D into two sets of components, $D = (D_{\text{obs}}, D_{\text{mis}})$, where D_{obs} denotes the constant components which are originally given (the observed data), and D_{mis} the missing components which are to be estimated by using D_{obs}. The missing data estimation task is to augment the missing values, i.e., assign values to elements in D_{mis}, in optimal manner with respect to the inference tasks for which the data is to be used. Here we restrict ourselves to *predictive inference tasks* (Bernardo & Smith, 1994; Gelman et al., 1995), i.e., we aim at developing augmentation methods that produce completions which result in good predictive performance, when the completed data is used to build a predictive model.

Completion criteria: Stochastic complexity

As discussed earlier, the approach adopted here is based on "scoring" alternative completions of the missing data D_{mis} based on an information-theoretic criterion. This criterion can be derived from the *Minimum Description Length (MDL) Principle* developed by Rissanen (1989,1996). According to MDL, the goal of all (inductive) inference from data is to *compress* the given data as much as possible, i.e., to describe it using as few bits as possible. Intuitively such an argumentation can be justified by the fact that to compress

BEST COPY AVAILABLE

data, one needs to find out regularities in it. The more we are able to compress the data, the more regularities (e.g., dependencies) we have found. Thus to be able to find the shortest encoding of data, we need to have extracted all the existing regularities. These regularities can then be used to characterize the underlying process generating the data, which is the purpose of modeling in the first place.

Data compression involves the use of a description method or *code*, which is a one–one mapping from datasets to their descriptions. Without loss of generality, these descriptions may be taken to be binary strings (Rissanen, 1989). Intuitively, the shorter the description or codelength of a set of D, the more regular or simpler the set D is. Rissanen (1987) defines the stochastic complexity informally as follows:

> The stochastic complexity of the data set D with respect to the model class $\mathcal{M}$ is the shortest code length of D obtainable when the encoding is done with the help of class $\mathcal{M}$ (Rissanen, 1987, 1996).

Here "with the help of" has a clear *intuitive* meaning: if there exists a model in $\mathcal{M}$ which captures the regularities in D well, or equivalently gives a good fit to D, then the code length of D should be short. However, it turns out to be very hard to define "with the help of" *in a formal manner*. Indeed, a completely satisfactory formal definition has only been found very recently (Rissanen, 1996).

Note that the informal definition of stochastic complexity (SC) as given above presumes the existence of a code: by definition, the SC of a data set D is the length of the encoding of D where the encoding is done using some special code C^* which gives "the shortest possible codelengths with respect to $\mathcal{M}$". In order to introduce a formula for the codelengths obtained using this C^*, the connection between probability distributions and codes has to be first clarified.

In general, we denote the length (in bits) of the encoding of D when the encoding is done using a code C by $L_C(D)$. All codes considered in MDL are *prefix* codes (Rissanen, 1989). From the Kraft inequality (see for example (Rissanen, 1989) or (Cover & Thomas, 1991)) it follows that for every prefix code C, there exists a corresponding probability distribution P such that for all data sets D of given length N (i.e., with N data instantiations), we have $-\log P(D) = L_C(D)$[1]. Similarly, for every probability distribution P defined over all data sets D of length N there exists a code C such that for all datasets D of length N, we have $L_C(D) = \lceil -\log P(D) \rceil$ (here $\lceil x \rceil$ is the smallest integer greater or equal to x). If we use $-\log P(D)$ instead of $\lceil -\log P(D) \rceil$, our code lengths will always be less than one bit off the mark; we may therefore safely neglect the integer requirement for code lengths (Rissanen, 1987). Once we have done this, the two facts above imply that we can interpret any probability distribution over sequences of a given length as a code and vice versa. This correspondence allows us to *identify* codes and probability distributions: every probability distribution P over data sets of length N may equivalently be interpreted as determining a code C such that $L_C(D) = -\log P(D)$ for all D of length N. We see that a short code length corresponds to a high probability and vice versa: whenever $P(D) > P(D')$, we have $-\log P(D) < -\log P(D')$.

If our parametric class of models $\mathcal{M}$ is regular enough (as it will indeed be for all instantiations of $\mathcal{M}$ we consider in this paper), then there exists a *maximum likelihood (ML) estimator* $\hat{\Theta}$ for every data set D, and we can write:

$$\begin{aligned} \hat{\Theta}(D) &= \arg\max_{\Theta \in \mathcal{M}} P(D|\Theta) \\ &= \arg\min_{\Theta \in \mathcal{M}} -\log P(D|\Theta) \\ &= \arg\min_{\Theta \in \mathcal{M}} L(D|\Theta), \end{aligned} \tag{1}$$

where the last equality indicates the fact that each Θ defines a code such that the code length of D is given by $-\log P(D|\Theta)$. Since this term can be interpreted as a code length, we denote it by $L(D|\Theta)$.

Let us now consider a data set D of arbitrary but fixed length N. The MDL Principle tells us to look for a short encoding of D. The model within class $\mathcal{M}$ that compresses the data most is the ML model $\hat{\Theta}(D)$, since by (1) it is the model for which $L(D|\Theta)$, the codelength of D when encoded with (the code corresponding to) Θ, is lowest. At first sight it seems that we should code our data D using $\hat{\Theta}(D)$, in which case the MDL Principle would reduce to the maximum likelihood method of classical statistics. However—and this is the crucial observation which makes MDL very different from maximum likelihood principle—MDL says that we must code our data using some *fixed* code, which compresses *all* data sets for which there is a good-fitting model in $\mathcal{M}$ (Rissanen, 1987). But the code corresponding to $\hat{\Theta}(D)$, i.e., the code that encodes any D' using $L(D'|\hat{\Theta}(D)) = -\log P(D'|\hat{\Theta}(D))$ bits, only gives optimal compression for *some* data sets (including D). For most other data sets $D' \neq D$, $\hat{\Theta}(D)$ will definitely not be optimal: if we had been given such a different data set D' (also of length N) instead of D, then the code corresponding to $\hat{\Theta}(D')$ rather than $\hat{\Theta}(D)$ would give us the optimal compression. In general, coding D' using $\hat{\Theta}(D)$ (i.e., using $L(D'|\hat{\Theta}(D))$ bits) may be very inefficient.

As discussed above, MDL says that we must code our data using some *fixed* code, which compresses *all* data sets that are well modeled by $\mathcal{M}$. We can therefore not use the code based on $\hat{\Theta}(D)$ if our data happens to be D and the code based on $\hat{\Theta}(D')$ if our data happens to be D': we would then encode D using a different code than when encoding D'. It would thus be very desirable if we could come up with a code that compresses each possible D as well as the maximum-likelihood, or equivalently, mostly-compressing element in

[1] Throughout this paper, by "log" we denote logarithm to the base two.

BEST COPY AVAILABLE

$\mathcal{M}$ for that specific D. In other words, we would like to have a single code C_1 such that $L_{C_1}(D) = L(D|\hat{\Theta}(D))$ for all possible D. However, such a code cannot exist as soon as our model class contains more than one element, since in general a code can only give short codelengths to a very limited number of data instantiations. Nevertheless, it is possible to construct a code C_2 such that

$$L_{C_2}(D) = -\log P(D|\hat{\Theta}(D)) + K_N = L(D|\hat{\Theta}(D)) + K_N \quad (2)$$

for all D of length N. Here K_N is a constant that may depend on N but is equal for all D of length N. If, for some $\Theta \in \mathcal{M}$, we say that it fits the data D well, we mean that the probability $P(D|\Theta)$ is high. Note that the code length obtained using C_2 precisely reflects for each D how well D is fitted by the model in the class that fits D best.

Picking C_2 such that the constant K_N is as small as possible yields the most efficient code that satisfies (2). We call the resulting code the *stochastic complexity code* and denote it by C^*. The corresponding minimal K_N is denoted by K_N^* and is called the *model cost* of $\mathcal{M}$. Consequently we are finally ready to give the formal definition of the stochastic complexity: the code length of D when encoded using the C^* code is the *stochastic complexity of D with respect to model class $\mathcal{M}$*, which we write as $SC(D|\mathcal{M})$:

$$\begin{aligned} SC(D|\mathcal{M}) &= L_{C^*}(D) \\ &= L(D|\hat{\Theta}(D)) + K_N^* \quad \text{where } \hat{\Theta}(D) \in \mathcal{M} \quad (3) \end{aligned}$$

In many situations the model classes are such that (3) cannot be easily calculated. Fortunately there exist several good approximations to SC (Rissanen, 1989, 1996). One good approximation is based on the equivalence between codes and probability distributions discussed earlier. As argued before, we can map code C^* to a probability distribution P^* such that for all D, $-\log P^*(D) = SC$. We saw that C^* can be seen as the code giving the shortest code lengths with respect to $\mathcal{M}$, similarly P^* can be seen as the probability distribution giving "as much probability as possible" to those data sets for which there is a good model in $\mathcal{M}$. A good candidate for such a distribution is the Bayesian *marginal likelihood* $P(D|\mathcal{M})$ (Bernardo & Smith, 1994), also sometimes known as the *evidence*, which can be computed by integrating over all possible models (parameter settings) Θ,

$$\begin{aligned} P(D|\mathcal{M}) &= P(D_{\text{obs}}, D_{\text{mis}}|\mathcal{M}) \\ &= \int P(D_{\text{obs}}, D_{\text{mis}}|\mathcal{M}, \Theta) P(\Theta|\mathcal{M})\, d\Theta. \quad (4) \end{aligned}$$

As discussed in (Rissanen, 1996) for many model classes (4) approximates $SC(D|\mathcal{M})$ extremely well, and thus in the sequel we will use this approximation as the "pragmatic" definition of stochastic complexity.

Stochastic complexity for Naive Bayes models

Since we are interested in categorical data, we use the set of (saturated) multinomial models for the complete data, which is a more general descriptive "language" than the multivariate normal class in the sense that it allows also for higher than two-way associations among the variables. In addition to selecting the multinomial model class some independence assumptions for the variables have to be made to make the stochastic complexity approach feasible in practice. This leads us to consider a special subclass of finite mixture models (Titterington et al., 1985) known as Naive Bayes models. For Naive Bayes model class, the categorical variables $X_i, i \neq s$ are assumed to be independent, given the values of a specific observed variable X_s, often called the *class variable*. For notational convenience we index these independent variables from $1, \ldots, m$. From this assumption it follows that the joint probability distribution for a data vector $\vec{d}$ can be written as

$$\begin{aligned} P(\vec{d}) &= P(X_1 = x_1, \ldots, X_m = x_m, X_s = k) \\ &= P(X_s = k) \prod_{i=1}^{m} P(X_i = x_i | X_s = k). \quad (5) \end{aligned}$$

Consequently, in the Naive Bayes model case, distribution $P(\vec{d})$ can be uniquely determined by fixing the values of the parameters $\Theta = (\alpha, \Phi)$,

$$\begin{aligned} \alpha &= (\alpha_1, \ldots, \alpha_K), \text{ and} \\ \Phi &= (\Phi_{11}, \ldots, \Phi_{1n}, \ldots, \Phi_{K1}, \ldots, \Phi_{Kn}), \end{aligned}$$

where the value of parameter α_k gives the probability

$$\begin{aligned} P(X_s &= k) \text{ and} \\ \Phi_{ki} &= (\phi_{ki1}, \ldots, \phi_{kin_i}), \\ \text{where } \phi_{kil} &= P(X_i = x_{il} | X_s = k). \end{aligned}$$

Here n_i denotes the number of possible values for variable X_i, and K the number of values for variable X_s. Using these denotations, we can now write

$$P(\vec{d}) = P(X_1 = x_{1l_1}, \ldots, X_m = x_{ml_m}, X_s = k) = \alpha_k \prod_{i=1}^{m} \phi_{kil_i}.$$

In the following we assume that $\alpha_k > 0$ and $\phi_{kil} > 0$ for all k, i, and l. Furthermore, both the variable distribution $P(X_s)$ and the conditional distributions $P(X_i|X_s = k)$ are multinomial, i.e., $X_s \sim \text{Multi}(1; \alpha_1, \ldots, \alpha_K)$, and $X_{i|k} \sim \text{Multi}(1; \phi_{ki1}, \ldots, \phi_{kin_i})$. Since the family of Dirichlet densities is *conjugate* (see e.g., (DeGroot, 1970)) to the family of multinomials, it is convenient to assume that the prior distributions of the parameters are from this family (see, e.g., (Heckerman, Geiger, & Chickering, 1995)). More precisely, let

$$\begin{aligned} (\alpha_1, \ldots, \alpha_K) &\sim \text{Di}(\mu_1, \ldots, \mu_K), \text{ and} \\ (\phi_{ki1}, \ldots, \phi_{kin_i}) &\sim \text{Di}(\sigma_{ki1}, \ldots, \sigma_{kin_i}), \end{aligned}$$

where $\{\mu_k, \sigma_{kil} \mid k = 1, \ldots, n_c; i = 1, \ldots, m; l = 1, \ldots, n_i\}$ are the *hyperparameters* of the corresponding distributions. For

Figure 1. The Expectation-Maximization algorithm for stochastic complexity minimization.

Algorithm 1
Expectation-Maximization (EM)

1. Set $t = 0$. Initialize parameters $\Theta^{(t)}$ randomly.
2. E-step: Determine

$$Q(\Theta, \Theta^{(t)}) = E[\log P(D_{\text{obs}}, D_{\text{mis}}|\Theta, \mathcal{M})P(\Theta|\mathcal{M})|D_{\text{obs}}, \Theta^{(t)}, \mathcal{M}],$$

where $\Theta^{(t)}$ are the parameter estimates in time step t.
3. M-step: Set $\Theta^{(t+1)} = \arg\max_{\Theta}\{Q(\Theta, \Theta^{(t)})|\Theta \in \Omega)\}$.
4. Set $t = t + 1$. Goto 2 if not converged.

Figure 2. The Simulated Annealing-algorithm for stochastic complexity minimization.

Algorithm 2
Simulated Annealing (SA)

1. Generate an initial random guess D_{mis} of the missing data. Set the *temperature parameter* T to its initial value.
2. Repeat L times

(a) Generate a new candidate $\tilde{D}_{\text{mis}}$ for the missing data by changing the estimate of one randomly chosen missing data item in D_{mis} to a randomly chosen value.

(b) If $P(D_{\text{obs}}, \tilde{D}_{\text{mis}}|\mathcal{M}) > P(D_{\text{obs}}, D_{\text{mis}}|\mathcal{M})$ then

$$\text{set } D_{\text{mis}} = \tilde{D}_{\text{mis}}.$$

(c) Else if $\left(\frac{P(D_{\text{obs}}, \tilde{D}_{\text{mis}}|\mathcal{M})}{P(D_{\text{obs}}, D_{\text{mis}}|\mathcal{M})}\right)^{1/T} > \text{Random}(0, 1)$ then

$$\text{set } D_{\text{mis}} = \tilde{D}_{\text{mis}}.$$

3. $T = F * T$. If not converged, goto 2.

simplicity, we will use here the uniform prior for the parameters, so all μ_k and σ_{kil} are set to 1. A more detailed discussion on the priors can be found in (Kontkanen, Myllymäki, Silander, Tirri, & Grünwald, 1998, 1997).

As shown in (Cooper & Herskovits, 1992; Heckerman et al., 1995), with the above assumptions the posterior probability of complete data $(D_{\text{obs}}, D_{\text{mis}})$ for a Naive Bayes model where n_c is the number of values for X_c is

$$\begin{aligned} & P(D_{\text{obs}}, D_{\text{mis}} \mid M_{n_c}) \\ = & \int P(D_{\text{obs}}, D_{\text{mis}} \mid \Theta, M_{n_c}) P(\Theta \mid M_{n_c})\, d\Theta \\ = & \frac{\Gamma\left(\sum_{k=1}^{n_c} \mu_{n_c}\right)}{\Gamma\left(N + \sum_{k=1}^{n_c} \mu_k\right)} \prod_{k=1}^{n_c} \frac{\Gamma(h_k + \mu_k)}{\Gamma(\mu_k)} \\ & \prod_{k=1}^{n_c} \prod_{i=1}^{m} \left(\frac{\Gamma\left(\sum_{l=1}^{n_i} \sigma_{kil}\right)}{\Gamma\left(h_k + \sum_{l=1}^{n_i} \sigma_{kil}\right)} \prod_{l=1}^{n_i} \frac{\Gamma(f_{kil} + \sigma_{kil})}{\Gamma(\sigma_{kil})} \right). \end{aligned} \tag{6}$$

Computing the stochastic complexity measure for the incomplete data matrix requires marginalizing out the missing data D_{mis}, i.e.,

$$\begin{aligned} SC(D_{\text{obs}} \mid M_{n_c}) &= -\log P(D_{\text{obs}} \mid M_{n_c}) \\ &= -\log \sum_{D_{\text{mis}}} P(D_{\text{obs}}, D_{\text{mis}} \mid M_{n_c}), \end{aligned}$$

where $P(D_{\text{obs}}, D_{\text{mis}} \mid M_{n_c})$ is given by (6), and the (exponential) sum goes over all the possible assignments of the missing data elements.

Search methods

Due to the exponential number of differenf completions of D_{mis}, for real data sets we cannot calculate SC for all possible completions. In general we have several alternative approaches for searching the data matrix completion with the minimal stochastic complexity. One possibility is to use a variant of the *Expectation-Maximization (EM)* algorithm (Dempster et al., 1977), which consists of two abstract steps, Expectation(E) and Maximization(M), presented in Figure 1.

One should observe that the EM algorithm does not provide an estimate of the missing data directly. The EM algorithm maximizes $P(\Theta|D_{\text{obs}}, \mathcal{M})$, and the resulting candidate for maximum posterior probability model $\hat{\Theta}$ can then be used for estimating the missing data D_{mis}. However, it can be shown that the stochastic complexity can be approximated by

$$SC(D_{\text{obs}}, D_{\text{mis}}|\mathcal{M}) \approx \log P(D_{\text{obs}}, D_{\text{mis}}|\hat{\Theta}, \mathcal{M})P(\hat{\Theta}|\mathcal{M}) - C, \tag{7}$$

where C is a constant depending on the number of the model parameters, and the number of the data vectors (Schwarz, 1978; Rissanen, 1989). As the expectation of the first term of this approximation is maximized during the EM process, it can be argued that the EM optimizes the stochastic complexity indirectly by optimizing the approximation (7).

An alternative is to use *simulated annealing (SA)* (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953; Kirkpatrick, Gelatt, & Vecchi, 1983), a stochastic global optimization method belonging to the family of Markov Chain Monte Carlo (MCMC) stochastic simulation algorithms. A commonly used version of SA goes is given in Figure 2. In this scheme, the *cooling factor* F is a constant parameter smaller than one. The SA algorithm converges as the temperature T approaches zero. It can be shown that if the initial temperature is high enough, and the decrement of the parameter is done slowly

BEST COPY AVAILABLE

enough, the process converges to the global optimum almost surely (Aarts & Korst, 1989).

Finally, by the *stochastic greedy (SG)* algorithm we mean a simple procedure where new solution candidates are generated as in simulated annealing, but the candidates are accepted only if the marginal likelihood is increased. Due to its simplicity and good performance, this stochastic greedy approach was used in the experimental part of the paper.

Experiments

Data

The use of the stochastic complexity based imputation will be illustrated using two different educational data sets containing categorical variables.

Subjects of the first "Popular kids" data set (POPKIDS) were students in grades 4-6 from three school districts in Ingham and Clinton Counties, Michigan. The data consists of 478 students were selected from urban, suburban, and rural school districts. In the collected questionnaires students indicated whether good grades, athletic ability, or popularity was most important to them. They also ranked four factors: grades, sports, looks, and money, in order of their importance for popularity. The questionnaire also asked for gender, grade level, and other demographic information. The study involved a classification task of correctly identifying the one of the six schools using the other variables as predictors.

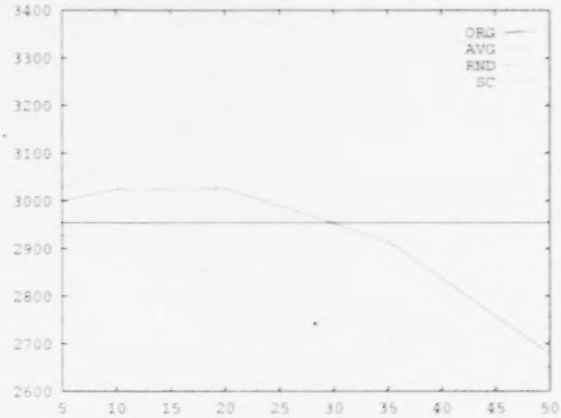
The second data set was "Irish educational transitions data" (IRISH) (Greaney & Kelleghan, 1984) reanalyzed by Raftery and Hout (1993). Subjects of this data set were 500 Irish schoolchildren aged 11 in 1967. The data were also used, in a simplified form, as an example to illustrate Bayesian model selection methods by Kass and Raftery (1994). The data had 6 variables, and the classification task was to predict the educational level (11 levels) of the students.

Experimental setting

In order to validate our approach, we produced synthetic missing data problems from the above real data sets by randomly deleting a known portion of the data. In the experiments also the sample sizes were controlled. Based on the different completions of D_{mis}, we performed a classification analysis, i.e., used the completed data sets to solve a classification problem in order to study the practical implications of different procedures for handling missing data.

For each data set, we created three different subsamples of sizes 10% 25% and 50% of the original data set size. Each time 50% of the data was reserved for the subsequent out-of-sample classification analysis. In each data sample we deleted (completely at random) 5%, 10%, 20%, 35% and 50% of the elements thus creating artificial missing data problems satisfying the missing completely at random

Figure 3. The value of the stochastic complexity measure (y-axis) of the imputed POPKIDS data matrix (N=239) as a function of the missing data percentage (x-axis). Different lines indicate the different imputation schemes (ORG=the original complete data matrix, AVG=imputing averages, RND=imputing by random, SC=imputing by minimizing the stochastic complexity).



(MCAR) assumption (Schafer, 1997). We then created complete data matrices by imputing the missing values using two alternative techniques. The first technique (AVG) was simply to impute the averages in the observed data. The second technique (SC) was to use a simple greedy algorithm where missing values were first imputed by random values (RND) and then one by one replaced by the values that minimize the stochastic complexity of the data. Since the missing data situation was artificially created, we were able to also use (as a reference point) the "oracle method" of always imputing the original value (ORG). These two techniques (AVG and SC) together with the two extreme reference techniques (RND and ORG) were then evaluated using both the stochastic complexity measure, and the percentage of correctly imputed values (correctness was judged by comparing the results to the original data). For each data set the setting described above was created 100 times and the averages were computed.

In order to study the practical implications of the different imputation schemes with different sample sizes and different percentages of missing data, all the imputations were used to build a model (classification rule) which was then used to classify previously unseen data items (train and test scheme). In this second set of experiments the result obtained by imputation schemes were further contrasted with the naive policy of building the model only using complete data items, i.e., those left intact by missing data generation. This policy corresponds to the case deletion methods widely used in practice. Again for each data set the setting described above was created 100 times and the averages were com-

BEST COPY AVAILABLE

Figure 4. The success rate (y-axis) of recovering the original complete POPKIDS data matrix (N=239) as a function of of the missing data percentage (x-axis). Different lines indicate the different imputation schemes (ORG=the original complete data matrix, AVG=imputing averages, RND=imputing by random, SC=imputing by minimizing the stochastic complexity).

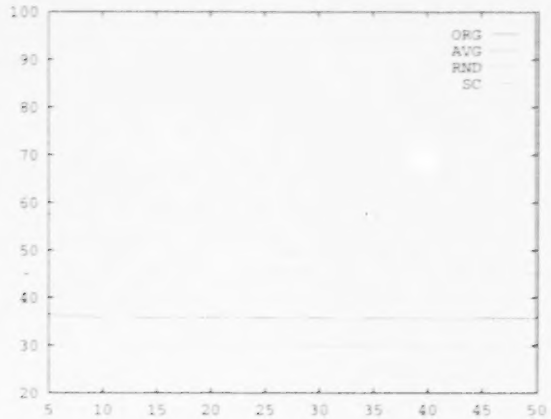
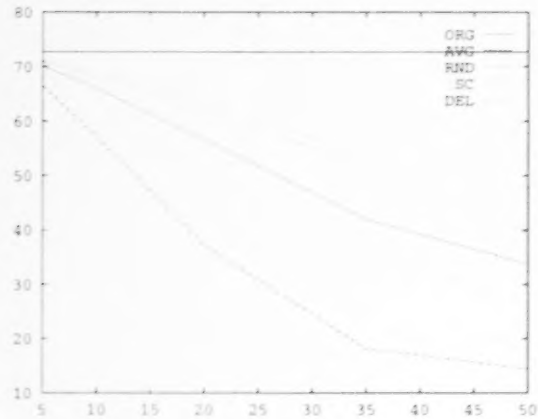


Figure 5. The classification accuracy (y-axis) as a function of of the missing data percentage (x-axis) when classifying 239 out-of-sample data vectors in the POPKIDS data set. Different lines indicate the performance of the models constructed from the complete data matrices obtained by different missing data handling schemes (ORG=the original complete data matrix, AVG=imputing averages, RND=imputing by random, SC=imputing by minimizing the stochastic complexity, DEL=case deletion).



puted.

Results

In Figure 3 we can see a typical example of the stochastic complexity measure for the completed data as a function of the different missing data percentages (POPKIDS data set). It is worth noticing that while only the SC-scheme attempts to minimize the stochastic complexity, also imputing sample averages (AVG) has the effect of minimizing complexity. When comparing the Figures 3 and 4 the percentage of correctly imputed values can be seen to follow the (negative of) stochastic complexity measure and is thus consistent with the theoretical analysis. In addition it should be observed that with respect to the average imputation approach the difference of recovering the original complete set is more than 15% on the average, and about 20% for missing data percentages less than 25%.

Comparing performance in classification also shows the beneficial effect of SC-based imputation (see Figure 5). The results also clearly demonstrate the inferior performance of the case deletion approach for prediction tasks, from pragmatic point of view the other commonly used technique, imputation of sample averages, is a significantly better alternative. However, imputing completely at random (RND) seems to yield a surprisingly good classification results. This is due to the fact that if missing completely at random assumption holds even approximately, imputation of random values does not bias the model construction.

The more detailed results for both data sets with all sample sizes, missing data percentages, and imputation methods are listed in Tables 1–6.

Summary and discussion

We have studied the problem of missing data estimation, and proposed a new information-theoretically justified approach to for incomplete multivariate categorical data. The approach discussed is a model-based imputation procedure relative to a model class, which in our case is the set of multinomial models with some independence assumptions. Based on the given model class assumption an information-theoretic criteria can be derived to select between the different complete data matrices. Thus the completion problem in this approach can be reduced to a search problem, i.e., finding the data completion with minimal criterion value.

As demonstrated by the empirical results, the stochastic complexity based approach performs well both in recovering the original missing data. More importantly, it also succeeds in augmenting the incomplete data matrix in such a manner, that in a subsequent classification task the complete data can be used to build a model that predicts better than the models built from complete data matrices produced by alternative methods. From practitioners point of view this latter aspect is more interesting, as completion of the incomplete data matrix is usually only an intermediate stage to applying various analysis tasks.

It should be observed that the approach adopted is very generic: changing the model class to a more descriptive one (e.g., to general graphical models such as Bayesian net-

BEST COPY AVAILABLE

Table 1

The value of the stochastic complexity measure of the imputed POPKIDS data matrix with different sample sizes as a function of the missing data percentage (M%). Different columns indicate the different imputation schemes (ORG=the original complete data matrix, AVG=imputing averages, RND=imputing by random, SC=imputing by minimizing the stochastic complexity).

Size	M%	AVG	RND	SC
47	5	638.83	640.72	631.23
	10	641.07	646.05	627.29
	20	639.57	653.97	620.56
	35	624.41	662.24	610.69
	50	588.79	667.62	599.39
119	5	1554.06	1563.32	1524.90
	10	1563.74	1587.13	1513.33
	20	1560.23	1623.03	1488.72
	35	1509.18	1661.02	1452.33
	50	1399.21	1684.87	1414.90
239	5	2999.84	3027.69	2928.01
	10	3025.04	3089.80	2899.97
	20	3026.34	3194.43	2841.77
	35	2915.83	3304.16	2747.40
	50	2679.93	3369.29	2649.96

Table 2

The success rate of recovering the original complete POPKIDS data matrix with different sample sizes as a function of the missing data percentage (M%). Different columns indicate the different imputation schemes (ORG=the original complete data matrix, AVG=imputing averages, RND=imputing by random, SC=imputing by minimizing the stochastic complexity).

Size	M%	AVG	RND	SC
47	5	34.76	29.12	53.04
	10	34.98	29.37	52.12
	20	35.32	29.95	47.55
	35	35.96	29.63	43.34
	50	35.55	29.60	38.22
119	5	35.68	29.88	57.34
	10	35.88	29.42	54.97
	20	35.87	29.92	51.68
	35	35.95	29.98	47.64
	50	36.12	29.83	42.14
239	5	36.50	29.95	57.67
	10	35.78	30.26	56.27
	20	35.89	29.89	54.02
	35	35.71	29.92	49.61
	50	35.78	29.67	45.06

Table 3

The classification accuracy as a function of of the missing data percentage (M%) when classifying 239 out-of-sample data vectors in the POPKIDS data set. Different columns indicate the performance of the models constructed from the complete data matrices (sizes 47, 119 and 239) obtained by different missing data handling schemes (ORG=the original complete data matrix, AVG=imputing averages, RND=imputing by random, SC=imputing by minimizing the stochastic complexity, DEL=case deletion).

Size	M%	AVG	RND	SC	DEL
47	5	46.39	46.87	46.03	41.68
	10	43.04	44.12	44.24	33.97
	20	34.81	39.82	39.85	24.03
	35	23.78	32.04	32.05	14.99
	50	17.75	23.75	25.07	14.23
119	5	60.67	61.15	61.13	55.21
	10	56.49	58.43	58.26	46.27
	20	46.00	52.30	52.85	31.10
	35	31.90	42.81	44.70	16.85
	50	25.58	30.10	33.82	14.38
239	5	70.08	70.86	71.13	66.47
	10	66.07	68.10	69.11	57.13
	20	56.59	61.83	63.29	37.21
	35	42.02	51.92	53.67	18.23
	50	33.81	39.60	42.93	14.45

Table 4

The value of the stochastic complexity measure of the imputed IRISH data matrix with different sample sizes as a function of the missing data percentage (M%). Different columns indicate the different imputation schemes (ORG=the original complete data matrix, AVG=imputing averages, RND=imputing by random, SC=imputing by minimizing the stochastic complexity).

Size	M%	AVG	RND	SC
50	5	380.38	381.33	373.58
	10	383.15	386.50	371.37
	20	384.56	394.21	367.10
	35	376.72	401.27	360.97
	50	355.28	406.76	355.29
125	5	901.80	907.08	877.15
	10	912.56	927.24	870.25
	20	917.87	958.75	856.93
	35	893.00	992.44	840.91
	50	829.52	1013.40	820.93
250	5	1729.24	1746.24	1670.96
	10	1755.35	1799.88	1656.73
	20	1765.85	1880.21	1627.49
	35	1708.64	1963.74	1584.08
	50	1572.05	2016.08	1538.59

BEST COPY AVAILABLE

Table 5

The success rate of recovering the original complete IRISH data matrix with different sample sizes as a function of the missing data percentage (M%). Different columns indicate the different imputation schemes (ORG=the original complete data matrix, AVG=imputing averages, RND=imputing by random, SC=imputing by minimizing the stochastic complexity).

Size	M%	AVG	RND	SC
	5	36.73	29.87	57.60
	10	36.47	29.80	57.23
50	20	35.90	31.72	54.82
	35	35.38	31.15	48.55
	50	34.78	29.80	42.52
	5	37.59	29.89	62.51
	10	37.56	29.81	59.37
125	20	37.05	30.49	56.79
	35	37.21	29.89	51.44
	50	36.73	30.34	44.85
	5	36.85	30.57	62.13
	10	37.08	30.03	60.37
250	20	37.58	30.49	57.44
	35	37.26	30.76	52.13
	50	37.11	30.56	46.69

Table 6

The classification accuracy as a function of of the missing data percentage (M%) when classifying 250 out-of-sample data vectors in the IRISH data set. Different columns indicate the performance of the models constructed from the complete data matrices (sizes 50, 125 and 250) obtained by different missing data handling schemes (ORG=the original complete data matrix, AVG=imputing averages, RND=imputing by random, SC=imputing by minimizing the stochastic complexity, DEL=case deletion).

Size	M%	AVG	RND	SC	DEL
	5	53.11	53.21	53.45	52.58
	10	50.30	51.58	52.04	50.08
50	20	43.14	49.19	50.30	44.56
	35	27.84	43.75	46.26	30.24
	50	17.72	33.69	38.06	10.79
	5	59.50	59.68	60.24	59.26
	10	57.26	58.66	59.35	57.05
125	20	49.02	55.40	57.31	50.34
	35	33.04	49.66	52.84	41.24
	50	20.70	41.18	44.80	22.62
	5	61.80	61.84	61.96	61.79
	10	60.34	61.47	61.80	61.06
250	20	53.80	59.86	60.74	56.91
	35	40.49	54.74	57.87	46.85
	50	27.01	48.19	51.40	30.80

BEST COPY AVAILABLE

works (Jensen, 1996; Pearl, 1988)) allows better "compression" of the data, i.e., better completions can be made. It is important to notice that in this sense the information-theoretic approach to modeling based on compression is akin to Bayesian modeling, which also always is relative to a set of models. Naturally missing data estimation is only one particular application of the MDL-approach, for other interesting applications such as model selection, time series analysis and predictive modeling the literature on minimum encoding modeling approaches should be consulted (see for example (Baxter & Oliver, 1994; Kontkanen, Myllymäki, Silander, & Tirri, 1997b; Rissanen, 1987, 1989; Wallace & Freeman, 1987) and references therein).

References

Aarts, E., & Korst, J. (1989). *Simulated annealing and Boltzmann machines: A stochastic approach to combinatorial optimization and neural computing*. Chichester: John Wiley & Sons.

Baxter, R., & Oliver, J. (1994). *MDL and MML: Similarities and differences* (Tech. Rep. No. 207). Department of Computer Science, Monash University.

Berger, J. (1985). *Statistical decision theory and bayesian analysis*. New York: Springer-Verlag.

Bernardo, J., & Smith, A. (1994). *Bayesian theory*. John Wiley.

Cooper, G., & Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, *9*, 309–347.

Cover, T., & Thomas, J. (1991). *Elements of information theory*. New York, NY: John Wiley & Sons.

DeGroot, M. (1970). *Optimal statistical decisions*. McGraw-Hill.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, *39*(1), 1–38.

Gelman, A., Carlin, J., Stern, H., & Rubin, D. (1995). *Bayesian data analysis*. Chapman & Hall.

Gilks, W. R., Richardson, S., & J., S. D. (1996). *Markov chain monte carlo in practice*. London, GB: Chapman & Hall.

Greaney, V., & Kelleghan, T. (Eds.). (1984). *Equality of opportunity in irish schools*. Dublin: Educational Company.

Heckerman, D., Geiger, D., & Chickering, D. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, *20*(3), 197–243.

Jensen, F. (1996). *An introduction to bayesian networks*. London: UCL Press.

Kass, R., & Raftery, A. (1994). *Bayes factors* (Tech. Rep. No. 254). Department of Statistics, University of Washington.

Kirkpatrick, S., Gelatt, D., & Vecchi, M. (1983). Optimization by simulated annealing. *Science*, *220*(4598), 671–680.

Kontkanen, P., Myllymäki, P., Silander, T., & Tirri, H. (1997a). Comparing stochastic complexity minimization algorithms in estimating missing data. In *Proceedings of wupes'97, the 4th workshop on uncertainty processing* (pp. 81–90). Prague, Czech Republic.

Kontkanen, P., Myllymäki, P., Silander, T., & Tirri, H. (1997b). On the accuracy of stochastic complexity approximations. In *Proceedings of the causal models and statistical learning seminar* (pp. 103–117). London, UK.

Kontkanen, P., Myllymaki, P., Silander, T., Tirri, H., & Grunwald, P. (1997). Comparing predictive inference methods for discrete domains. In *Proceedings of the sixth international workshop on artificial intelligence and statistics* (pp. 311–318). Ft. Lauderdale, Florida.

Kontkanen, P., Myllymaki, P., Silander, T., Tirri, H., & Grunwald, P. (1998). Bayesian and information-theoretic priors for Bayesian network parameters. In *Proceedings of the 10th european conference on machine learning (ECML-98)*. Chemnitz, Germany. ((To appear).)

Little, R., & Rubin, D. (1987). *Statistical analysis with missing data*. Wiley.

McLachlan, G., & Thriyambakam, K. (Eds.). (1997). *The EM algorithm and extensions*. New York: John Wiley & Sons.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, M., & Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chem. Phys., 21*, 1087–1092.

O'Hagan, A. (1994). *Kendall's advanced theory of statistics. volume 2b: Bayesian inference*. Cambridge: Edward Arnold.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann Publishers, San Mateo, CA.

Raftery, A., & Hout, M. (1993). Maximally maintained inequality: Expansion, reform and opportunity in irish schools. *Sociology of Education, 66*, 41–62.

Rissanen, J. (1984). Universal coding, information, prediction, and estimation. *IEEE Trans. on Inf. Theory, IT-30*(4), 629–636.

Rissanen, J. (1987). Stochastic complexity. *Journal of the Royal Statistical Society, 49*(3), 223–239 and 252–265.

Rissanen, J. (1989). *Stochastic complexity in statistical inquiry*. New Jersey: World Scientific Publishing Company.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory, 42*(1), 40–47.

Rubin, D. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons.

Rubin, D. (1996). Multiple inputation after 18 years. *Journal of the American Statistical Association, 91*, 473–478.

Schafer, J. (1995). Model-based imputation of census short-form items. In *Proceedings of the annual research conference* (pp. 267–299). Washington, DC: Bureau of the Census.

Schafer, J. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461–464.

Titterington, D., Smith, A., & Makov, U. (1985). *Statistical analysis of finite mixture distributions*. New York: John Wiley & Sons.

Wallace, C., & Freeman, P. (1987). Estimation and inference by compact coding. *Journal of the Royal Statistical Society, 49*(3), 240–265.

BEST COPY AVAILABLE